Meeting WG Log – Hadoop for OHDSI

Attendees:

* Anthony Reckard – QuintilesIMS
* Malcolm McRoberts – NantHealth
* Ke Wang – Quintiles IMS
* Lee Evans – LTS Computing
* Tom White – Cloudera
* Melissa Crenshaw – NantHealth
* Brack Coalson – Cloudera
* Siva – Quintiles IMS
* Chetan – Quintiles IMS
* Sajeev Cherian – Abbvie
* Vartika Singh - Cloudera
* Chris Wallace – Quintiles IMS
* Cornelius Raths – Nanthealth
* Sanjeev Garr - Abbvie
* NOT IN ATTENDANCE: Georgia Tech, Takeda

New People

* Sajeev – leads pharmacovigilance and safety, Abbvie, includes adverse events, big data, signal detection.

New Business

* F2F – Georgia Tech.
* Hadoop Breakout. Friday 17th. Georgia Tech Research Institute.
  + Attendees: No one?! Chetan may go.
  + What You Want From Breakout: Topics of interest?
    - Prospective Topics: What is Hadoop (technical); Why Hadoop for RWE (sort of technical); Odysseus on state of progress to spark based conversions; GA Tech
  + Who wants to demo: no one from the call.

Other

* NantHealth: sort of stuck on migrating \_\_\_\_ to Impala. Does not have Apache Kudu (which has read/write), but Apache Impala.
  + - From Tom: Cloudera has GA Kudu (for a few weeks). No work done with Kudu for OMOP. Why Kudu or not Kudu? (in Tom’s humble opinion): Kudu is storage, Impala is query engine, so either Impala on HDFS or Impala on Kudu. Some cases you won’t have to update data frequently, was either non-update or batch update (not individual records).
    - Nant: was thinking about Achilles for this – wanted to test Kudu with Achilles. Struggling to get things to run in Impala. Specifically ability of Impala to support Varchar(8) to represent dates (vs. Varchar(10)).
    - Tom on ‘types’ or ‘data types’ in Impala. For Achilles (not OMOP), changes to SQL Render or some component to accommodate Impala, affected another db (Postgres) so had to step back. Need to figure out how to support date types in Impala. Currently using varchar.
    - Perhaps warrants a deep tech dive on this outside of this forum.
    - Shawn will send a note a public service announcement on Achilles not working on Impala.
    - Other packages Nant wants to get working on Impala: Feature extraction project.
    - Do people want to use Hadoop/Impala other underneath just OMOP or all the OHDSI projects?
    - Chetan: They want to leverage multiple projects and tools in OHDSI, not just OMOP!
    - Lee: a fundamental question. Lee deploys OHDSI tools, (and modifies SQL render for unsupported platforms i.e. Netezza). Large pharma companies generally want everything: their data converted into CDM, and they want the tools a) ATLAS is the main tool, includes R, b) Achilles.
    - Nant: All

Call for questions: send survey

Other topics

* Chetan: some other packages (e.g. patient level prediction) work off of ‘R’. Is there a simple way to support R in Hadoop? (e.g. SparkR or SparklyR). How does that work?
  + Vartika: a blog is out on how to set up SparklyR (Vartika recommends SparklyR in terms of feature/function and adoption by our customers). SparkML resolves perhaps some issues with data types outside of Hadoop.
  + Tom: or Spark-ML-Lib perhaps. Ad hoc discovery based science use of R is one thing; the production part of using R is kind of a different thing. SparkMLLib is better for the second.
  + Malcolm: played with R packages – if you start greenfield, SparkR looks like R. But the data types are different, so *you can’t engage existing R libraries*.
  + Vartika: SparkR on its own does not parallelize well. SparklyR has deep bindings. More people moving to Python to SparkR? (for scaling purposes)
* Chetan: if Chetan downloads the other OHDSI packages, will they run with no additional work? (Dolley thinks there will be more work). Shawn will document (with Chetan if he has) some of the steps and work needed to get one of the other workgroup packages
* Chetan: Cloudera VM to get started (can we build a VM or a Docker), that has packages already installed (Tom knows what this is).
* Lee: manages OHDSI cloud, and Atlas projects, and the databases. OHDSI cloud could have Kudu and/or Impala running. Could we do this through a PMI grant to Columbia? Yes, however AWS Infrastructure costs may be an issue.
* Sanjeev: Abbvie starting to use R and Hadoop, Sanjeev has a technical team member or peer. Interested in how other organizations are using it.